



GeT
Génome et
Transcriptome

Toulouse Genomic Core facility

Olivier Bouchez

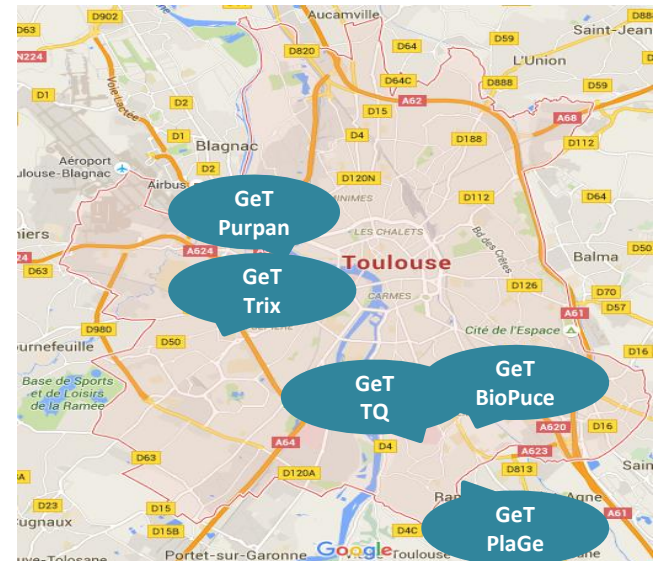
Next Generation Sequencing Lab Manager

<http://get.genotoul.fr>

@GeT_Genotoul

Who are we?

- **Genomics and transcriptomics core facility spread on 5 sites GeT in Toulouse**
- **National Infrastructure within the « France Génomique » program**
- **IBISA Label**
- **INRA strategic core-facility**
- **ISO9001 et NFX50-900 Certification**



Team and Expertise

- A team of 30 people with:
 - Technological Specialty and Scientific community by site
 - Experts in Agronomy, Environment, Microbiology, Health
 - Competence in biology, bioinformatics, biostatistics



- Partnership with Genotoul Bioinformatic core-facility (NG6) for:
 - Data storage and management
 - Data Quality analysis

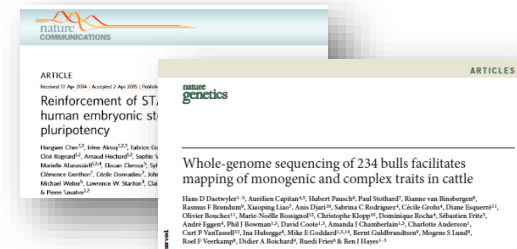
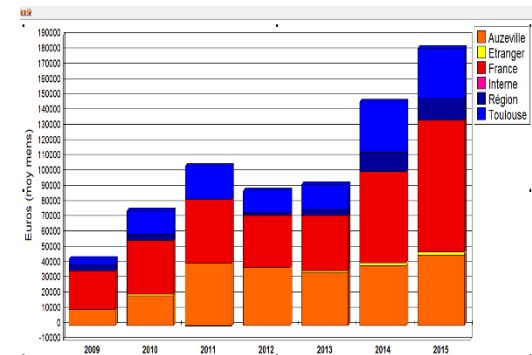


The Core facility missions

- **To provide innovating technologies for genome analysis to the scientific community**
 - Sequencing / Genotyping
 - Gene expression
- **To Develop new protocols, new methodologies, acquire expertise and train in those technologies**
- **To animate workshop for user network**



- **15 R&D projects**
- **More than 100 laboratories (INRA, CNRS, INSA, INSERM, CHU, CIRAD ...)**
 - More than 160 research teams
 - More than 250 projects
 - 3M€ of activity
- **17 Research projects (ANR, INCA, H2020 ...)**
- **46 Publications as co-author since 2012**



Technologies

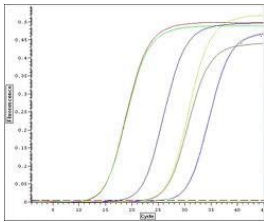
Tools to improve the activity

- **Sample and library quality controls**
- **Single cell capture (C1 Fluidigm)**
- **Pipetting platforms for sample preparation**
 - Partnership with Tecan (4 Evo), Agilent Bravo
 - Access array (fluidigm)



Tools to analyse gene expression and to genotype

PCR quantitative



**ViiA7, QuantStudio,
ABI7900HT, ...**



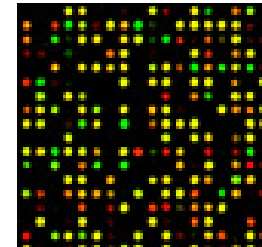
PCR quantitative microfluidic



BioMark (Fluidigm)



Microarray



**Affymetrix – Agilent
iScan**



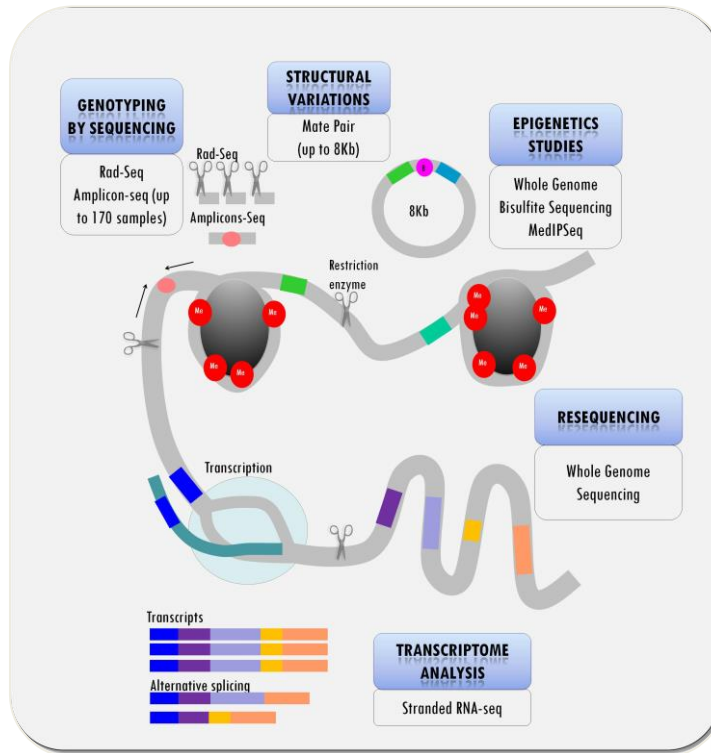
From Sanger to short read NGS revolution



400 pb
1Gb



200 pb
13 Gb



2x 300 pb
15 Gb



2x150 pb
700 Gb

From 1 human genome sequencing to....

... « 1000 genomes » projects for all species

From NGS short reads to long reads

- **First PACBIO installed in France**
- **SUNRISE project to acquire expertise**
 - **To validate quality of DNA**
 - **To improve library preparations**
 - **To increase the number of reads**
 - **To increase the length of reads**



IGM (San Diego, USA) 202 SMR cells		NUM	MAX	N50 BP	N50 NUM	MEAN	MEDIAN	BP/SMRTcell
	moyenne	98666	45457	12211	28413	9176	9032	0,906 M
	max	146374	52725	12981	41602	9997	9809	1,36Gb
Lausanne University (Swiss) 59 SMRT cells								
	moyenne	106800	46800	15172	28371	10773	9821	1,15Gb
	max	144358	53253	16132	38325	11436	10568	1,6GB
Get-PlaGe (France) 146 SMRT cells								
	moyenne	77086,6301	52317,4932	15365,4795	19705,5822	10326,6773	9152,5137	800Mb
	max	126777	80974	20507	33133	13635	12295	1,3Gb

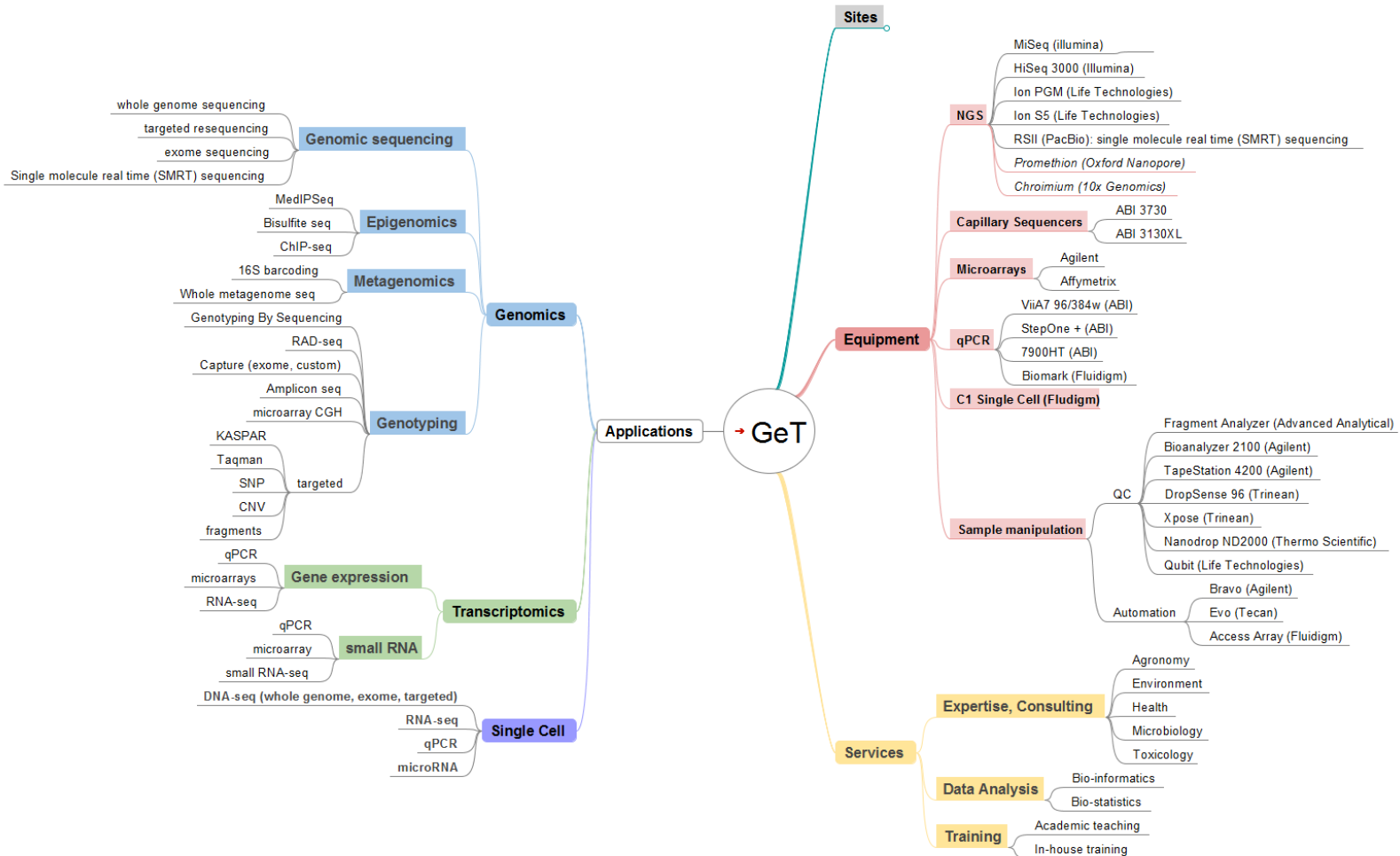
Improvements of the molecular biology steps have increased the length of the Pacio Sequences (B. Mayjonade)

From NGS short reads to long reads

- **Fifteen projects to develop new applications :**
 - **Whole genome sequencing on different species**
 - **Targeted sequencing**
 - **Complex population & metagenomics**
 - **RNA sequencing**
 - **Epigenetic**



GeT in a mapmind





GET
Génome et
Transcriptome

Focus on Next Generation Sequencing

From Sanger to 3rd generation

Unique sequence

Next Generation Sequencing

Sanger



16 or 48 capillaries

NGS - short reads

Ion Torrent

Ion PGM
 2 Gb
 400 pb



Ion S5
 15 Gb
 200 pb

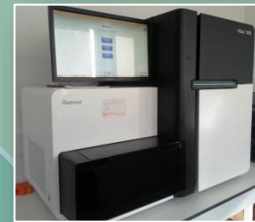


3 x MiSeq
 15 Gb
 2 x 300 pb



Illumina

2xHiSeq 3000
 700 Gb
 2 x 150 pb



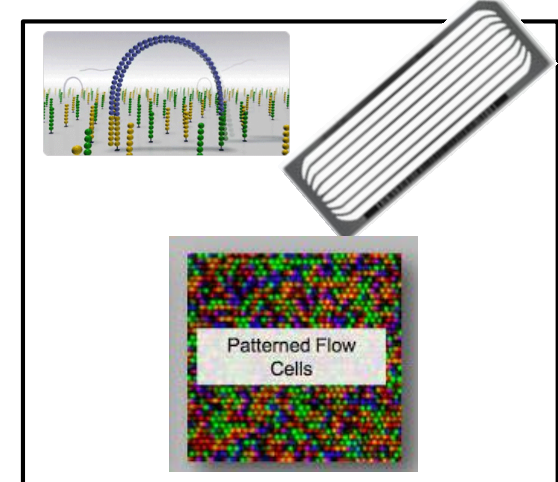
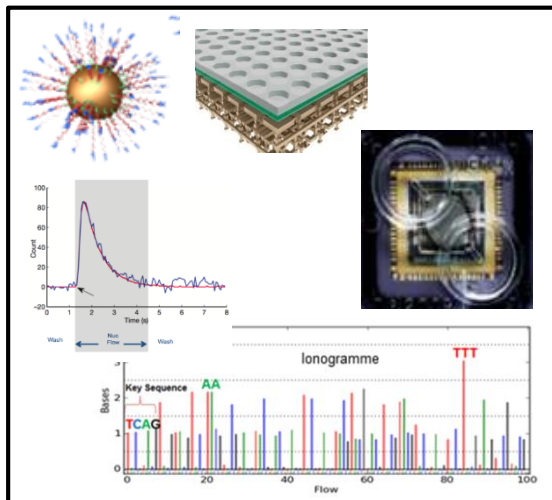
3G - long reads
PacBio RslI
 70 000 reads - 20 kb



NGS : Ion Torrent et Illumina



Ion Torrent	Illumina
Scalable: different chips ↔ output: 50 Mb to 15 Gb	Output: 2 Gb to 90 Gb / lane
Reads : 1 x 200 pb or 1 x 400 pb	Reads: 1 x 50 pb to 2 x 300 pb
Bioinformatic tools pre-defined	Many développements ongoing Data stored on the bioinformatic core facility's clusters
New: S5 sequencer	New: 2nd HiSeq3000



NGS: which applications ?

Genomics Re-Sequencing

- **Whole Genome**
- **Whole Exome**
- **Gene panels**
- **Hotspots**

Transcritomics Expression

- **Whole transcriptome**
- **Full Length Transcripts**
- **Small RNA**
- **Targeted RNAseq**

Epigenetics

- **Whole Genome**
- **Targeted**

Metagenomics

- **16S**
- **Amplicons**

Genomics – Re-sequencing/ *de novo* sequencing

➤ Whole Genome

Available:

- "PCR-free" librairies for Illumina
- 1 HiSeq3000 run = 8 genomes at 30X \approx 2,5 days
- Ion S5 for small genomes
- GBS

⇒ mutations screening

Other applications :

- Sequencing on PacBio RSII: 100 SMRT cells \Leftrightarrow 3 Gb genome at 30X
- Sequencing on PacBio RSII: 1 SMRT cell \Leftrightarrow 1 bacterial genome at 100X

⇒ *De novo* sequencing, screening for chromosome modifications, haplotype detection, SNP identification...

Illumina HiSeq 3000



PacBio RslI



Ion S5



Transcriptomics - Expression

➤ Whole transcriptome sequencing

⇒ Identification / quantification

Available:

Stranded RNAseq (HiSeq3000, 2 x 150 pb)

➤ Full length transcripts sequencing

Under development on PacBio RsII

⇒ Fusion transcripts, alternative transcripts

➤ small RNA sequencing

Available on Ion S5 sequencing

< 100 pb – stranded sequencing

ILLUMINA HiSeq



PacBio RsII



ION S5



Transcriptomics - Expression

➤ Targeted RNAseq ⇒ Expression

**Design possible for custom panels:
On Ion Torrent and Illumina**

➤ Targeted RNAseq - gene fusion

On Ion Torrent and Illumina :

- RNA apoptose / RNA cancer panel
- TruSight RNA Pan-Cancer...
- custom

Ion PGM



Ion S5



Illumina MiSeq



➤ Bisulfite conversion

Available for whole genome on HiSeq
RRBS (Reduced Representation Bisulfite Sequencing)

➤ lncRNA/small RNA

RNAseq protocols modifications (ribodepletion vs poly-A selection)

➤ ChIP-seq/MedIP-seq

Illumina compatible

➤ "direct" methylome

Available on PacBio RslI for small genomes

Illumina HiSeq



Ion S5



PacBio RslI



Metagenomics - Metabarcoding

➤ 16S – commercial kit

- 7 variable regions on 9
- Analysis on Ion Reporter
- Low throughput

➤ 16S full length

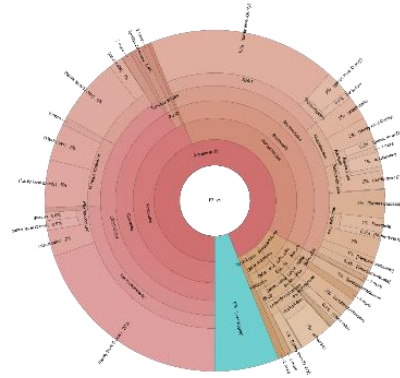
- Available PacBio RsII

➤ Custom Amplicons

- Available on Illumina – fully automated
- > 300 barcodes available
- 1 region 300-400 pb

➤ Whole metagenome / transcriptome

- Available on HiSeq



Ion PGM



PacBio RsII



Illumina MiSeq



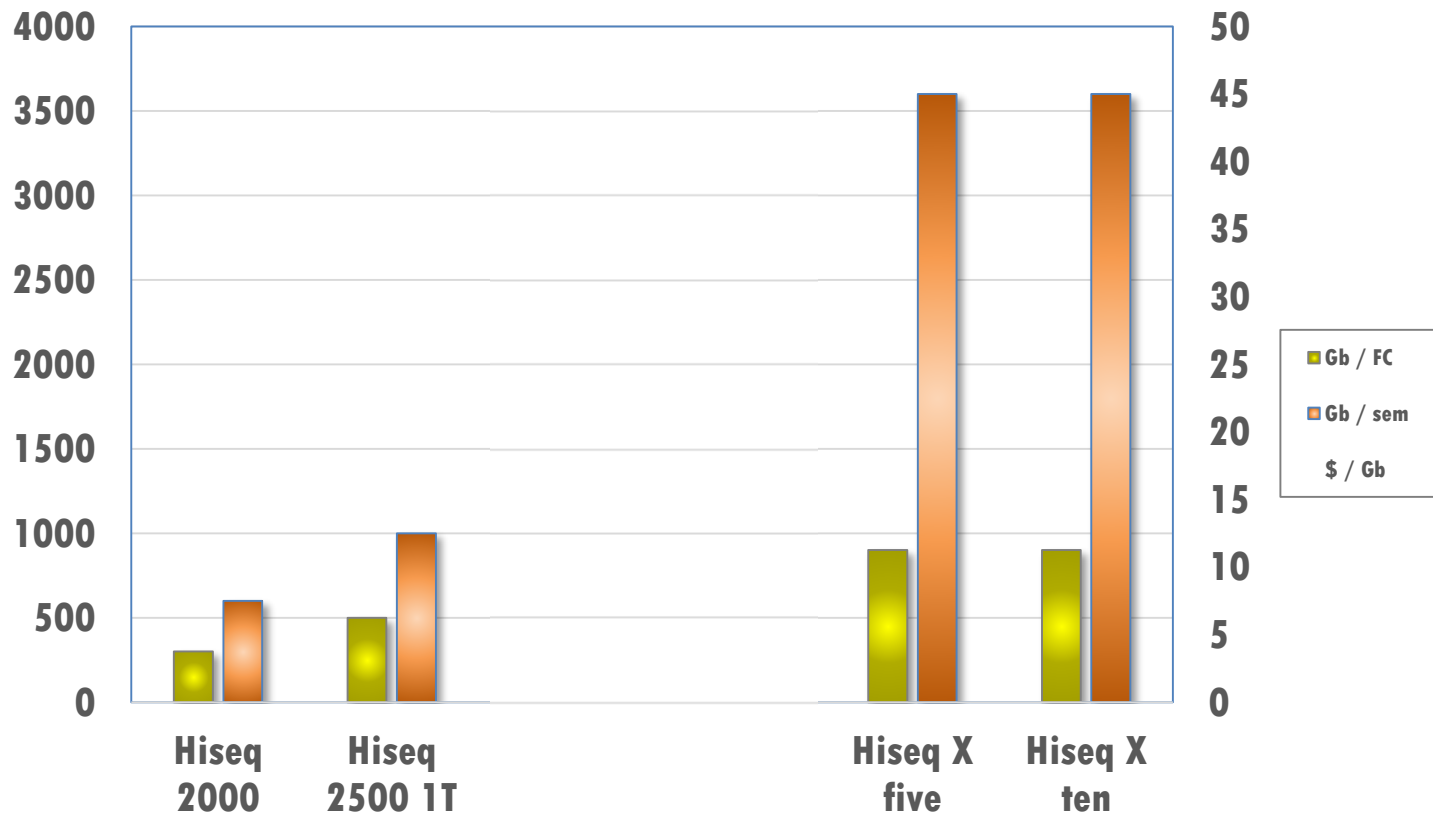
Ion S5



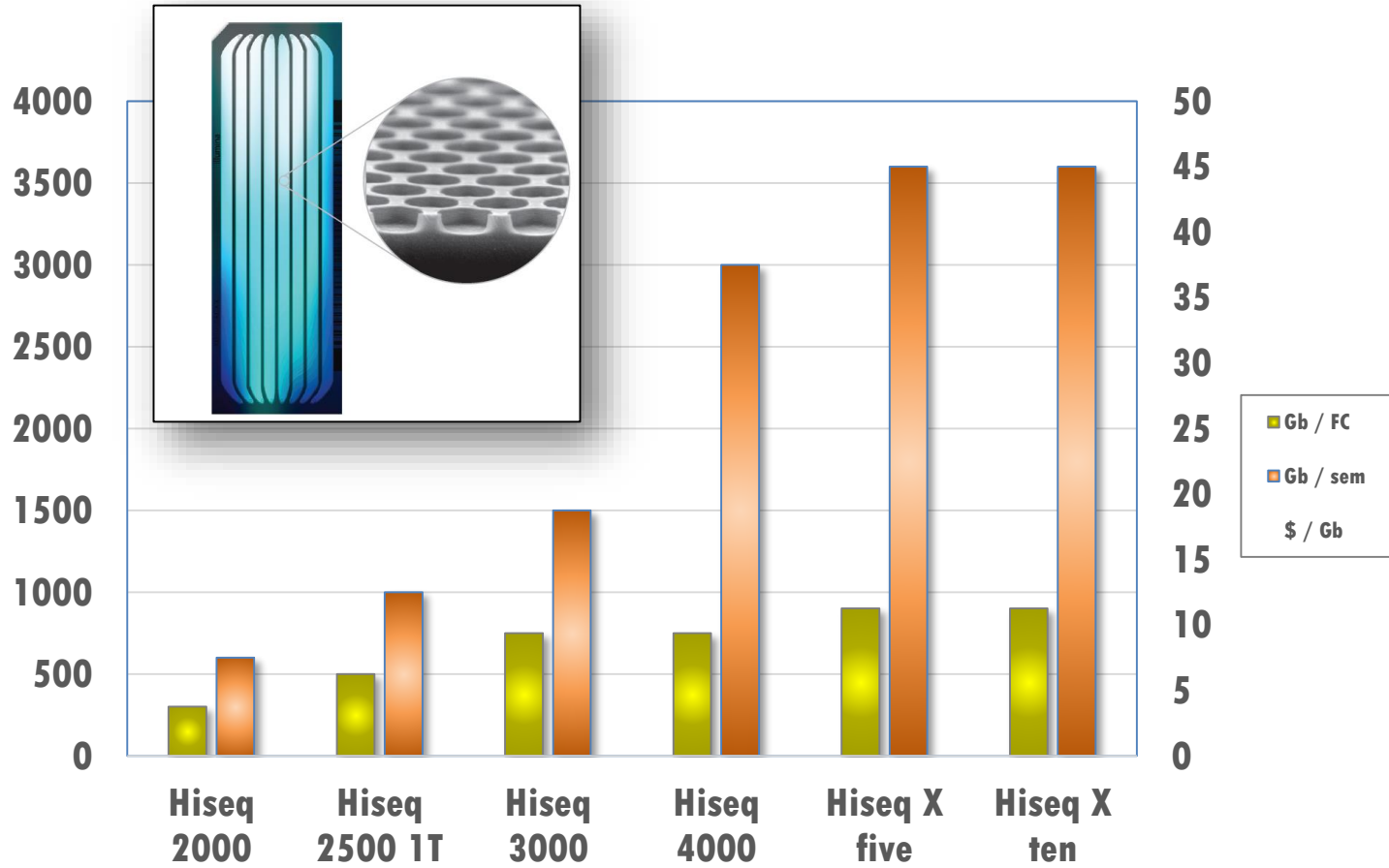
Illumina HiSeq



HiSeq productivity and costs



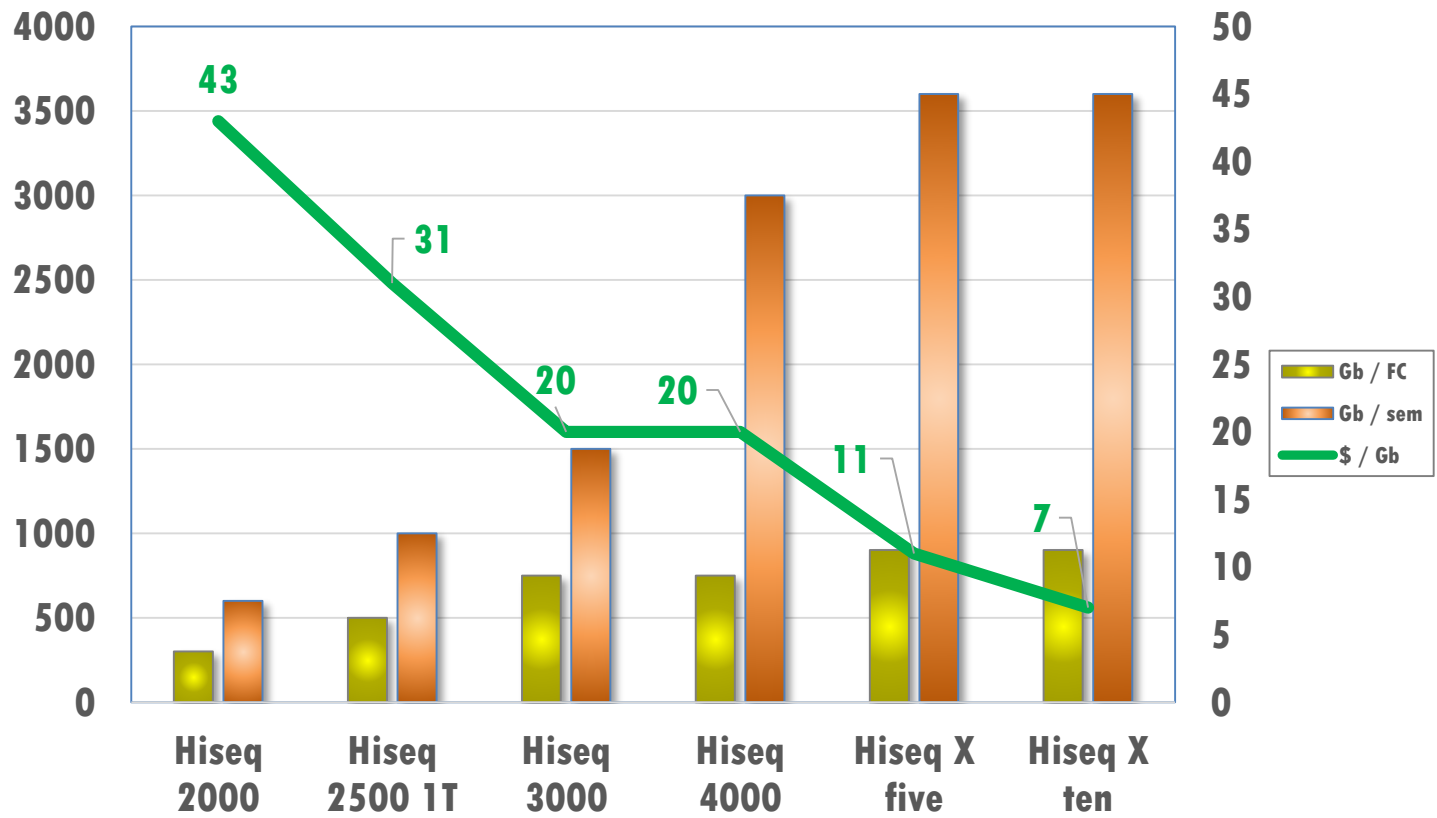
HiSeq productivity and costs



January 2015 : announcement of HiSeq3000 & HiSeq4000
 (same Flow Cells as HiSeq X)

HiSeq productivity and costs

ILLUMINA REAGENT COSTS \$ / Gb



October 2015 : HiSeqX open to non human (30x coverage) but 5 HiSeqX at least...



GET
Génome et
Transcriptome

Example of PacBio sequencing projects

3rd generation: PacBio RsII

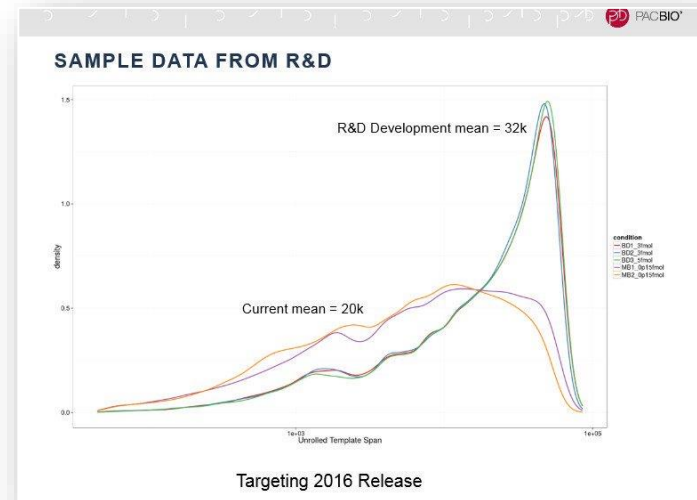
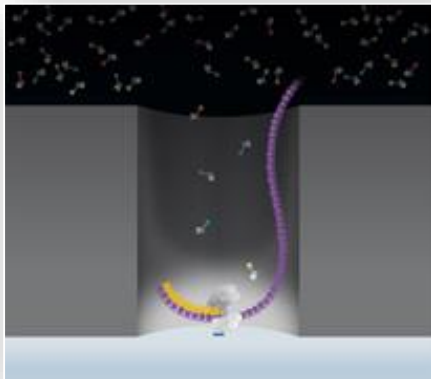


- + Single molecule real time sequencing
- + No PCR
- + Reads: 15-20 kb \Rightarrow 40 kb

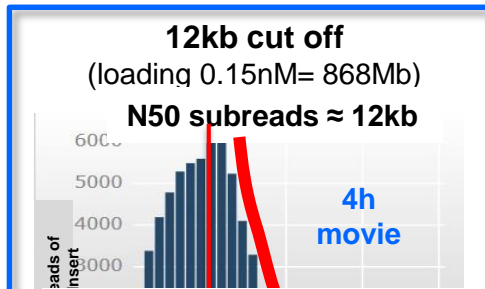
- High error rate (>10%)



Random errors => corrected thanks to sequencing depth



Read length increase



IGM (San Diego, USA) 202 SMRT cells

	NUM	MAX	N50 BP	N50 NUM	MEAN	MEDIAN	BP/SMRTcell
moyenne	98666	45457	12211	28413	9176	9032	0,906 M
max	146374	52725	12981	41602	9997	9809	1,36Gb

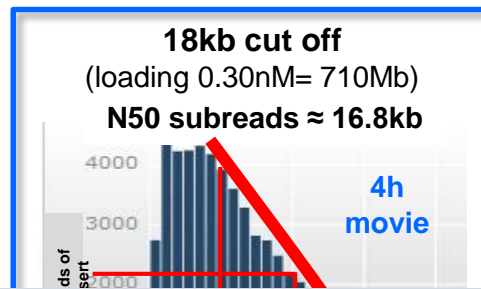
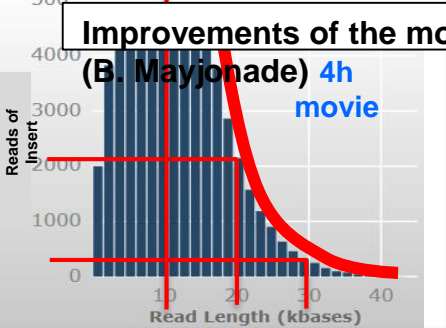
Lausanne University (Swiss) 59 SMRT cells

moyenne	106800	46800	15172	28371	10773	9821	1,15Gb
max	144358	53253	16132	38325	11436	10568	1,6Gb

Get-PlaGe (France) 146 SMRT cells

moyenne	77086,6301	52317,4932	15365,4795	19705,5822	10326,6773	9152,5137	800Mb
max	126777	80974	20507	33133	13635	12295	1,3Gb

N50 subreads ≈ 15kD



IGM (San Diego, USA) 202 SMRT cells

	NUM	MAX	N50 BP	N50 NUM	MEAN	MEDIAN	BP/SMRTcell
moyenne	98666	45457	12211	28413	9176	9032	0,906 M
max	146374	52725	12981	41602	9997	9809	1,36Gb

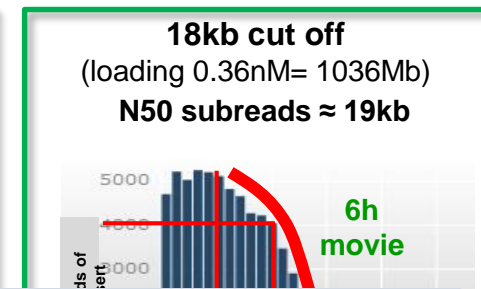
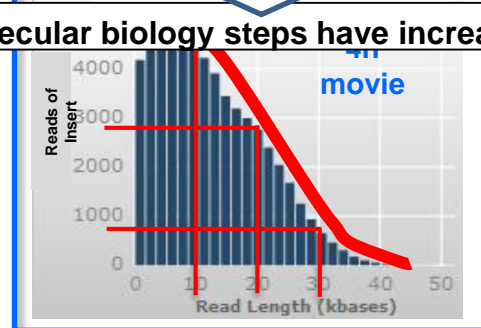
Lausanne University (Swiss) 59 SMRT cells

moyenne	106800	46800	15172	28371	10773	9821	1,15Gb
max	144358	53253	16132	38325	11436	10568	1,6Gb

Get-PlaGe (France) 146 SMRT cells

moyenne	77086,6301	52317,4932	15365,4795	19705,5822	10326,6773	9152,5137	800Mb
max	126777	80974	20507	33133	13635	12295	1,3Gb

N50 subreads ≈ 17.5kD



IGM (San Diego, USA) 202 SMRT cells

	NUM	MAX	N50 BP	N50 NUM	MEAN	MEDIAN	BP/SMRTcell
moyenne	98666	45457	12211	28413	9176	9032	0,906 M
max	146374	52725	12981	41602	9997	9809	1,36Gb

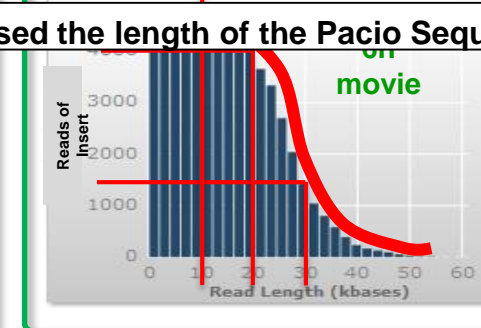
Lausanne University (Swiss) 59 SMRT cells

moyenne	106800	46800	15172	28371	10773	9821	1,15Gb
max	144358	53253	16132	38325	11436	10568	1,6Gb

Get-PlaGe (France) 146 SMRT cells

moyenne	77086,6301	52317,4932	15365,4795	19705,5822	10326,6773	9152,5137	800Mb
max	126777	80974	20507	33133	13635	12295	1,3Gb

N50 subreads ≈ 20.5kD



Improvements of the molecular biology steps have increased the length of the Pacio Sequences

(B. Mayjonade) 4h movie

Read length increase

147 SMRT on 1st sunflower genome:

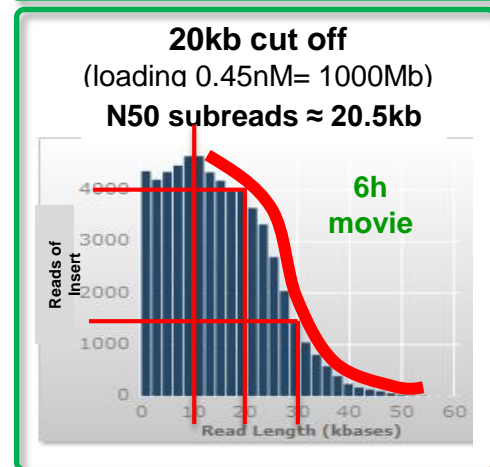
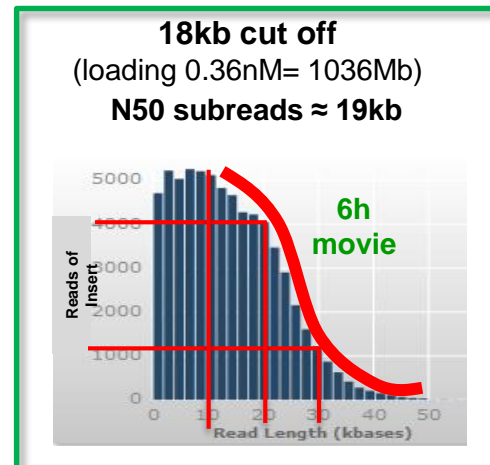
- **N50 15365**
- **800 Mb / SMRT cell**

103 SMRT on 2nd sunflower genome:

- **N50 18510**
- **1 041 Mb / SMRT cell (max 1 445 Mb)**

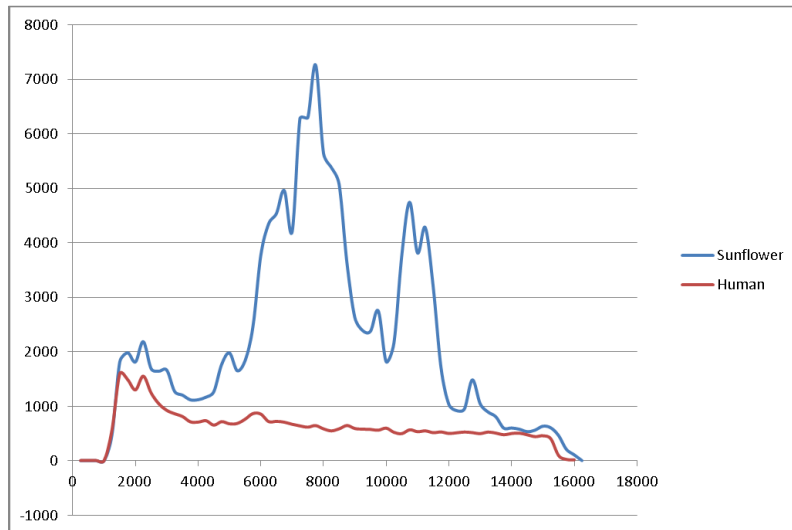
Top 10 of our longest subreads

80974 bp
 79860 bp
 79834 bp
 78105 bp
 77481 bp
 76881 bp
 76558 bp
 76355 bp
 75569 bp
 75559 bp



Sunflower genome sequencing

Sunflower : 30 % repeated sequences(LTR)
Homme : 8.8 % repeated sequences



The Challenge: obtention of sequences longer than the repeats

- INRA (Sunflower team):**
 Hiseq, 127 X
 → **43 % genome covered**
- International consortium:**
 454, Hiseq, Genetic & Physic maps (BAC)
 → **63 % genome covered**
- INRA (Sunflower team & GeT-PlaGe) :**
 PacBio, 107 X (407 SMRT)
 → **84 % genome covered**

#ctg	MAX	N50 BP	# > N50	MEDIAN	Gb
13 124	4.4M	498 kb	1700	118 kb	3.03

Direct methylation analysis on PacBio

Goal: identification of different methylation patterns on evolved strains of the bacteria *Ralstonia solanacearum* from PacBio sequencing

2 strains: ancestral strain & evolved strain

3 SMRTcells/strain

Direct methylation identified from PacBio sequencing:

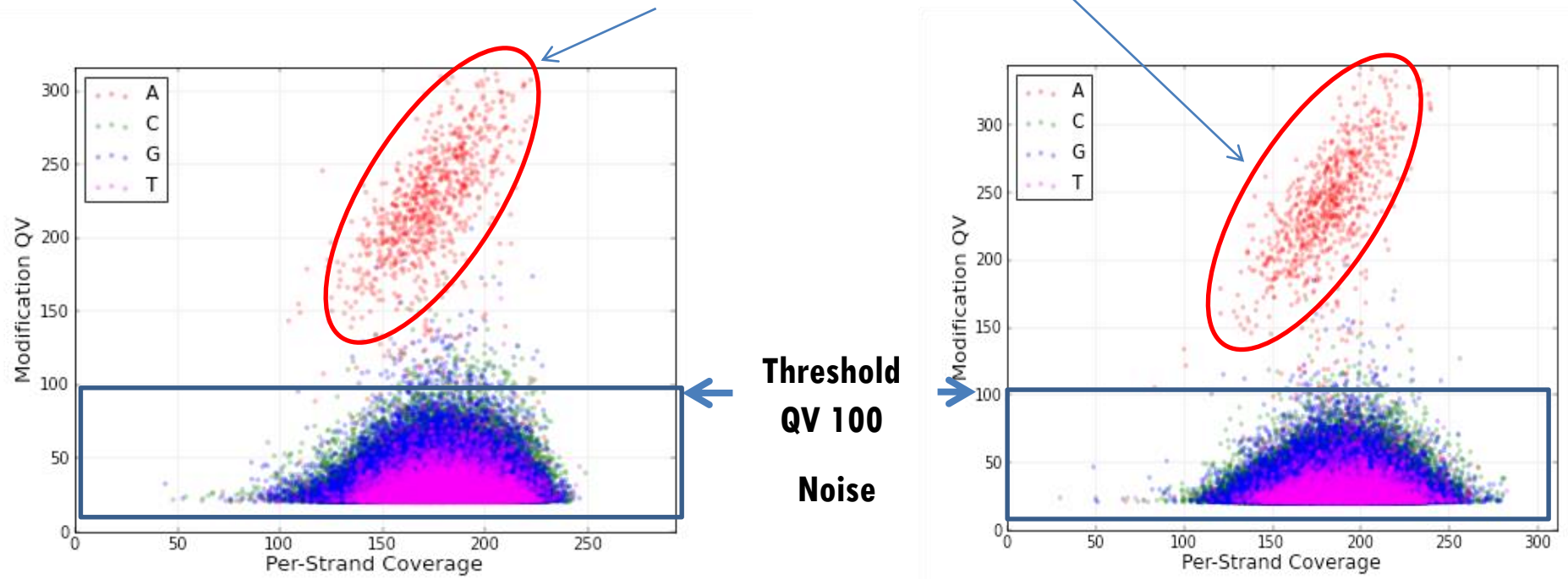
- ✓ 6mA @50X
- ✓ 4mC @50X
- ✓ 5mC @500X

Modifications identified

Ancestral strain

Evolved strain

6mA methylation signal



Preliminary results

6mA pattern: GTWWAC

784 GTWWAC patterns in the genome

>98% of 6mA with a méthylation at a QV >100

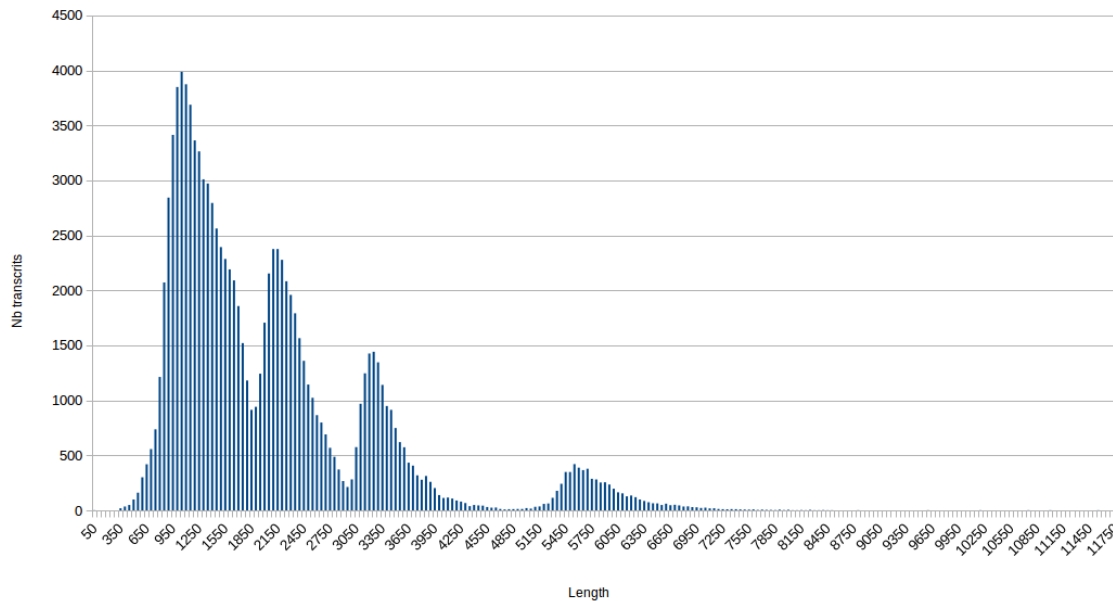
>98% 6mA -> typical pattern for bacteria

Preliminary analysis:

- ✓ **GTWWAC pattern associated with 99,2% of the 6mA**
- ✓ **GCCGGC pattern associated with 33.2% of the 4mC**
- ✓ **5mC difficile à voir en PacBio**

~100 different patterns identified between ancestral and evolved strains => genes or intergenic regions, including many regulatory genes

Iso-seq = full length transcripts sequencing on PacBio RSII



Preliminary results : ~110000 different transcripts identified



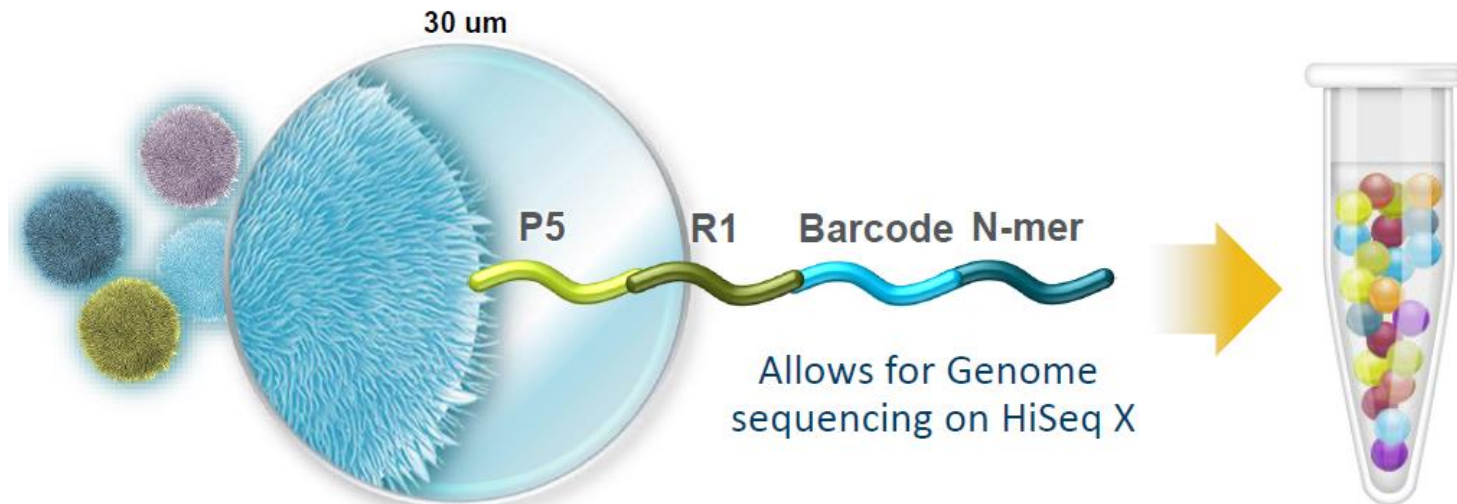
GET
Génome et
Transcriptome

New applications

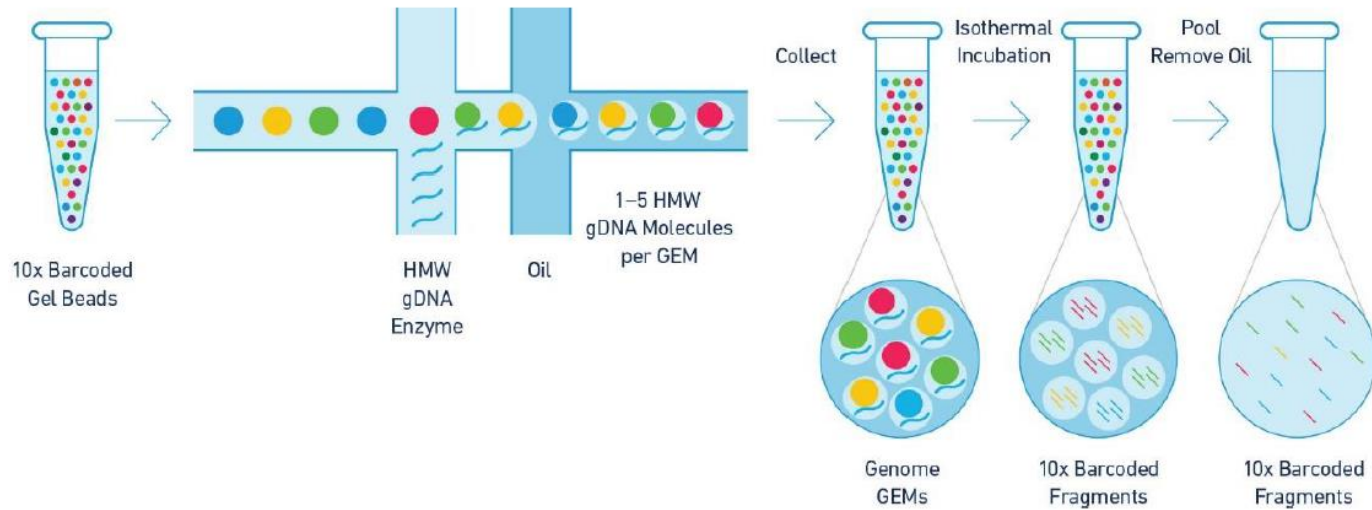
New technology available

Chromium (10X genomics)

- Library prep for Illumina sequencing
- Long range genomic (~100 kb), haplotyping
- Single cell analysis
- Exome sequencing



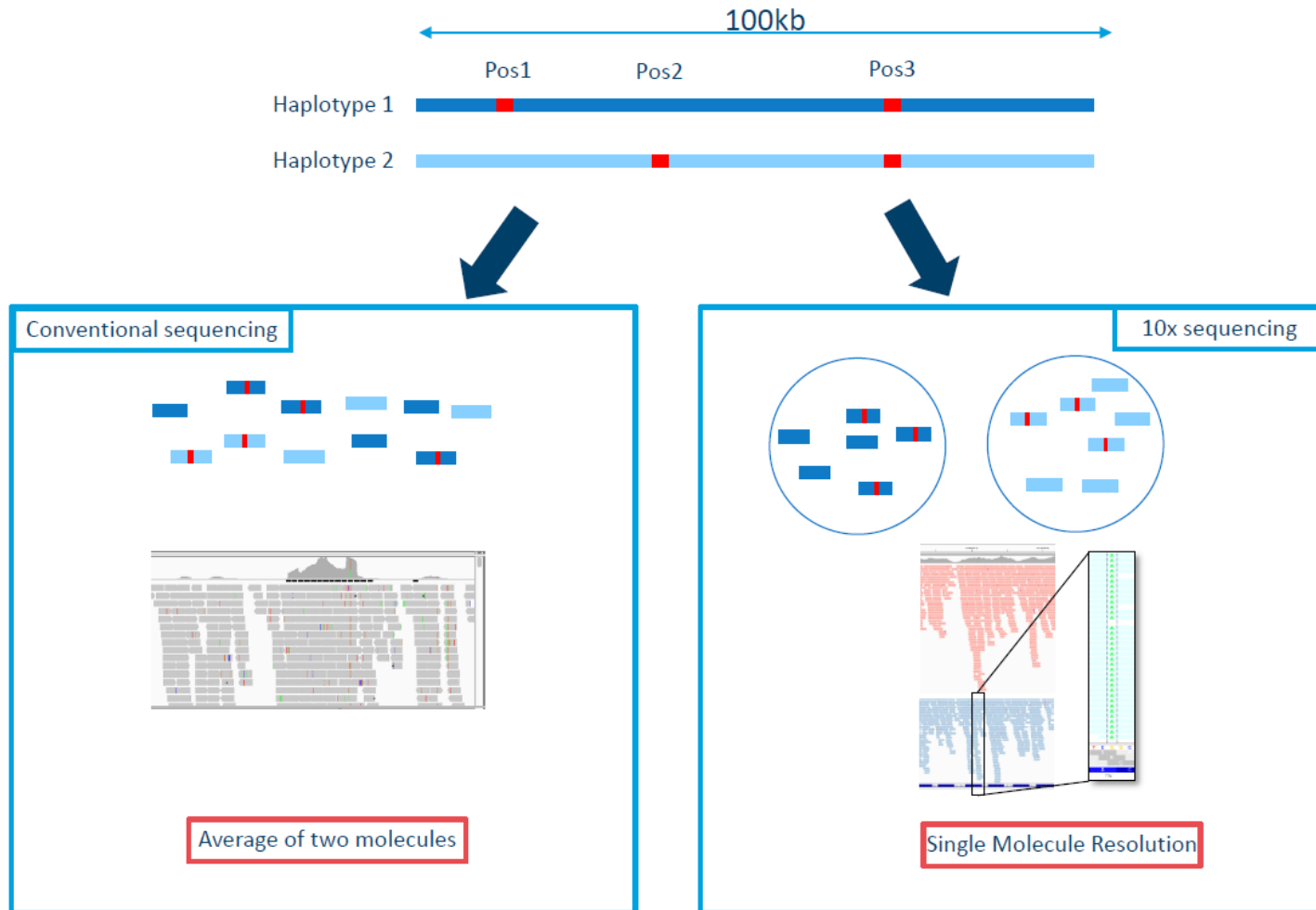
10X GENOMICS



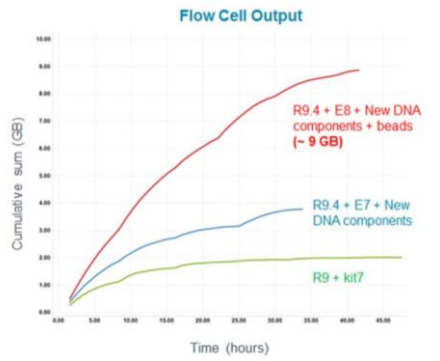
Linked-Reads



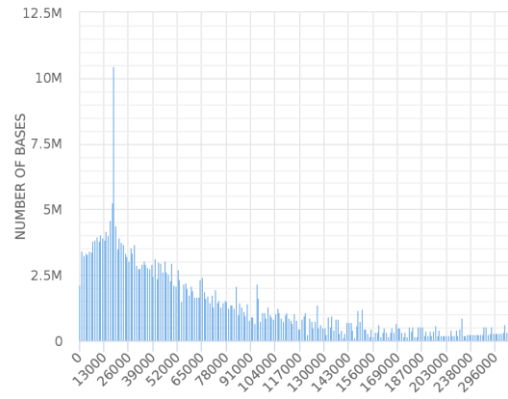
10X GENOMICS



Oxford Nanopore



BASES SEQUENCED BY READ LENGTH



MinION



smidgION

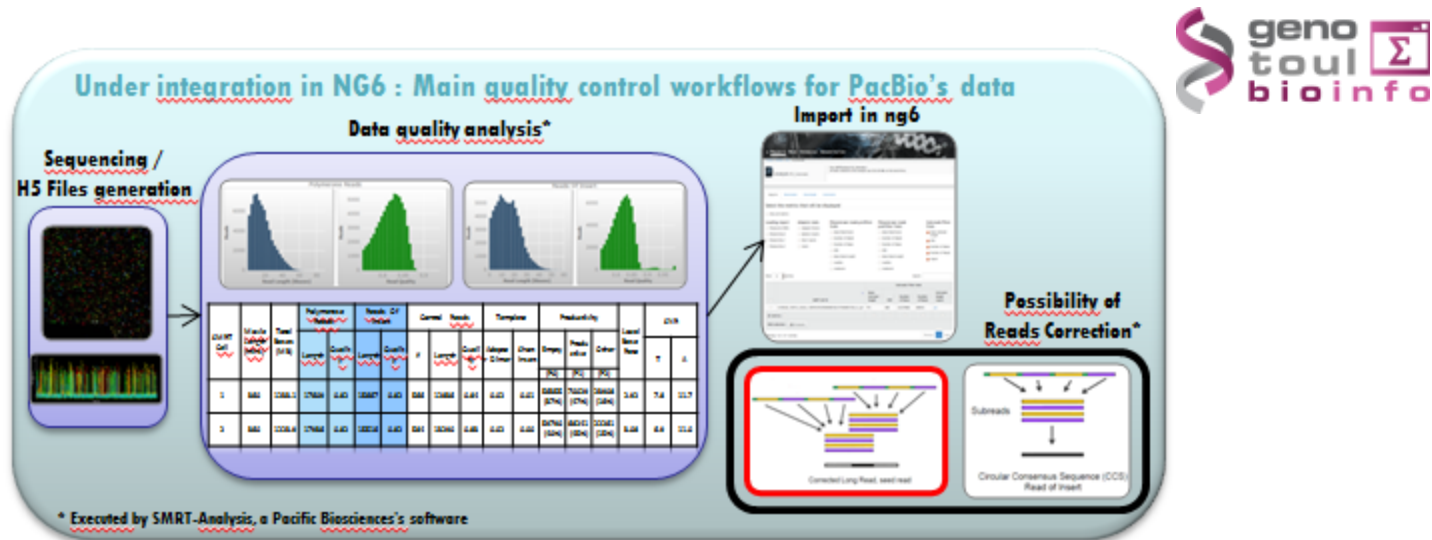


PromethION



LIMS development

- **Current integration in NG6 : Main quality control workflows for PacBio's data**



- **Upcoming : A new LIMS for NGS samples, sequencing and analysis tracking**



The future?



GOUVERNEMENT.fr
 22 juin 2016 - Communiqué
Remise du rapport "France Médecine Génomique 2025" : faire entrer la France dans l'ère de la médecine génomique



Illumina HiSeq X ten : 18000 genomes/year, 1000\$ genome



genomeweb
 Business & Policy Technology Research Clinical Disease Areas Applied Markets Resources
 Home » Business, Policy & Funding » Research Funding » France Plans to Invest €670M in Genomics, Personalized Medicine
France Plans to Invest €670M in Genomics, Personalized Medicine
 Jun 23, 2016 | a GenomeWeb staff reporter
 NEW YORK (GenomeWeb) – The French government announced that it plans to invest €670 million (\$760.8 million) in a genomics and personalized medicine program meant to improve the diagnosis and prevention of disease in the country.

**Target: 230000 genomes/year (2020)
12 sites in France**

Création Unité de Service GeT-PlaGe



Création d'une entité INRA Transfert hébergée sur la plateforme : services aux entreprises

Vers une feuille de route des infrastructures de recherche de l'INRA



GET

Génome et
Transcriptome

Thanks!